systems. To identify tax evasion among registered firms, DGT takes steps to determine the accuracy of tax returns, including implementing a compliance risk management model that predicts the likelihood of non-compliance. These predictions are partly based on tax audit results covering around 2 percent of registered firms each year. However, there is no public reporting on the scale of misreporting identified by these audits, nor any estimates of total tax losses if rates of identified misreporting were to be extrapolated.

3 Data

The data used in this analysis comes from the 2023 WBES in Indonesia, which interviewed the top managers or owners of 2,955 firms. This sample was nationally representative of firms possessing a Company Registration Certificate (TDP) or Business Identification Number (NIB) with five or more employees with at least 1 percent of private ownership and who do not have legal status as cooperatives. Firms are selected through stratified random sampling from the 2016 Economic Census conducted by the Central Agency of Statistics of Indonesia. Stratification is based on sector, firm size (employment),[7] and location. The sectoral distribution covers firms from the manufacturing and major services sectors. The manufacturing sector includes all firms with activity defined under Section C of ISIC Rev 4.0. In contrast, the services sectors include Section F (Construction), Section G (Wholesale and Retail Trade), Section H (Transport and storage), Section I (Accommodation and food service activities), Section M (Professional, technical, and scientific services), as well as the following two-digit codes: 58 (Publishing activities), 61 (Telecommunications), 62 (Computer programming), 79 (Travel agencies), 95 (Repairs).

The survey covers all 38 provinces of Indonesia, though some are aggregated, resulting in a total of 22 regional strata. The sample design ensures no more than 7.5 percent margins of error and 90 percent confidence intervals at each of the stratification levels: size, sector,

[7]The stratification by firm size is defined using the number of employees in the firm. The definitions are consistent across all countries where the WBES has been implemented and are defined as small (5-19 employees), medium (20 to 99 employees), and large (100 and more).

and region. All interviews were conducted face-to-face with top managers or business owners in Bahasa Indonesian through Computer-Assisted Personal Interviews (CAPI) using tablets and Survey Solutions as the software for data collection. Data collection started in December 2022 and concluded in September 2023, achieving a response rate of 41.2 percent, similar to response rates for WBESs in other countries. The replacement of businesses refusing to participate in the survey was done within the same stratum, following a randomly generated preference order of which businesses should be contacted for an interview. All interviews were subject to the WBES protocols[8] of quality assurance. Prior to the launch of data collection, the team of interviewers underwent a 5-day training to understand the methodology, protocols, and each question that appears in the instrument. The training was used to verify and improve the instrument's translation, and this was fine-tuned through pilot interviews with a small sample of firms before the formal launch of data collection.

The survey primarily consisted of detailed questions about firm characteristics and their operating environment, followed by the double list experiment (see the following section). The questions about firm characteristics and their operating environment are part of standardized modules included in WBESs that have been implemented in more than 150 countries. This includes specific questions about how firms have interacted with the government (e.g., the time required to obtain permits and encounters with corruption) and firms' experience paying tax (e.g., which types of tax they must pay and if tax officials have visited them). The extensive list of questions about firm characteristics and their operating environment provides a rich dataset to examine the heterogeneity of the double list experiment.

4 Design of the Experiment

A "double" list experiment was included in the 2023 WBES in Indonesia to estimate levels of tax evasion by firms. List experiments (also known as the item count technique) provide a way to estimate levels of tax evasion without having to directly ask respondents if they

[8]Details about the methodology and protocols can be found here: WBES Methodology

engage in such activities. This indirect approach is typically seen as a more credible way of asking questions about sensitive topics as respondents provide "concealed" answers that can still be used to generate estimates of sensitive responses (originally proposed and elaborated on by Raghavarao and Federer, 1979; Miller, 1984; Droitcour et al., 2004). Specifically, list experiments involve randomly allocating respondents into two groups. They are then shown a list of statements and asked to state the number of statements (but not which specific statements) are true in their case. The list shown to the "treatment" group includes one extra statement about the sensitive issue (e.g., not paying all the taxes that they owe). While it is not known which specific respondents have agreed with the statement on the sensitive issue, differences in the average number of statements reported as true between the treatment and control groups can be used to estimate the prevalence of the sensitive response. If respondents do not engage in sensitive activities, the mean number of items should be the same between the control and treatment groups. The double list experiment approach is an extension of the traditional single list experiment: respondents complete two list experiments sequentially and are randomly allocated to the treatment group in one of the two.

The double list experiment included as part of the WBES is shown in Table 1. Respondents were randomly assigned to either Group A or Group B. Each group was given two list questions to respond to, but with a slight variation between the groups. Group A received the treatment in List 2 while being the control for List 1 (note the statement about taxes in Group A's List 2). Group B had the opposite arrangement, receiving the treatment in List 1 while serving as control in List 2. The treatment is the inclusion of a sensitive item asking about tax evasion, specifically "This establishment does not pay all the taxes it is required to pay." The exact location of this statement in the list of statements was randomized. This text is identical to what was used in previous studies asking directly about tax evasion and has been trialed as part of single list experiments on firms in other middle-income countries (e.g., see Dom et al., 2022). The other statements in the double

list experiment were carefully determined to minimize the risk that respondents could gauge the focus of the experiments. Consultations with private sector experts in Indonesia were conducted to ensure the statements achieve this aim, and they were subsequently fine-tuned through extensive piloting. In the case of List 1 in Table 1, the first statement is expected to apply to most respondents, while the second item is expected to be very uncommon. In the case of List 2, the second item (and the third item to a lesser extent) was expected to apply to most respondents, while the first item was expected to be much less common.

Table 1: Design of Experiment

	Group A	Group B
List 1	a) This establishment had to let go of an employee over the last year b) This establishment's last month sales increased by 200% c) The establishment was temporarily closed during the COVID-19 pandemic	a) This establishment had to let go of an employee over the last year b) This establishment's last month sales increased by 200% c) The establishment was temporarily closed during the COVID-19 pandemic **d) This establishment does not pay all the taxes it is required to pay**
List 2	a) This establishment almost went bankrupt in the last year b) At least one of the employees of this establishment contracted COVID-19 since the start of the pandemic c) The price of the main product of this establishment changed over the past year **d) This establishment does not pay all the taxes it is required to pay**	a) This establishment almost went bankrupt in the last year b) At least one of the employees of this establishment contracted COVID-19 since the start of the pandemic c) The price of the main product of this establishment changed over the past year

The experiment was administered during the interview with the top manager or owner of the firm, following the survey section about interactions with the government. The assignment to group A or B was randomly determined when the interview appointment was scheduled (see the balance table, Table A1 in the Appendix, showing that random assignment was successful). The interviewers were not informed whether the respondent was assigned to group A or B. Once the interview reached the point in which the experiment was administered, the interviewers informed the respondent that they were only interested in the number of items that were accurate for them, emphasizing that they were not interested in which ones were accurate and that they could not comment on or provide any clarifications of any

of the items. The interviewers showed respondents the list of items to count on the tablet on which survey responses were being collected. The CAPI system randomized the order in which the items appeared. Respondents only provided the number of accurate items in each list.

The double list experiment was designed with several considerations in mind based on an extensive review of best practices in the recent literature (e.g., Imai, 2011; Blair and Imai, 2012; Glynn, 2013; Rosenfeld et al., 2016; Ahlquist, 2018; Chuang et al., 2021). Firstly, the number of items provided to respondents must be brief to minimize respondent fatigue but long enough to reassure respondents that their answers will be concealed credibly. As such, we chose three "non-sensitive" items alongside the tax evasion item, which is consistent with many recent studies. Secondly, it is essential to avoid "ceiling" and "zero" effects (where most respondents report none of or all the items on the list). Otherwise, it will be evident to the respondent that their answer to the tax evasion item will not be concealed. Therefore, we included one non-sensitive item most respondents were expected to select and another most respondents were not likely to. Thirdly, the theme of the non-sensitive items should be broadly related to the sensitive item. Otherwise, the latter may stand out too starkly to respondents, and they may be less inclined to report honestly. To address this, we ensure all non-sensitive items in both list experiments are framed around the business activities of the "establishment" that the respondent is answering on behalf of. Finally, we field a double list experiment with many strengths relative to a traditional single list experiment, including the fact that answers can be compared between the two list experiments to see if a similar prevalence of tax evasion is reported. By doing so, we provide credible evidence to dispel one of the most common criticisms of list experiments: that the non-sensitive items on the list may drive respondents' willingness to answer honestly. A double list experiment allows us to check whether similar levels of tax evasion are reported across the two lists with different non-sensitive items.

5 Findings

5.1 Main Results

Figure 1 shows the distribution of responses to the two list questions, disaggregated by treatment (blue) and control (pink) groups. The horizontal axis represents the number of accurate statements in the list experiments selected by respondents. The treatment group was presented with a list of four statements, while the control group was presented with a list of three statements. Most respondents had selected that just one or two statements were accurate in both lists of experiments. Relatively few selected all or none of the statements (which provides strong evidence against ceiling or floor effects being a significant concern). Both list experiments demonstrated that respondents allocated to the treatment group (i.e., Group B in List 1 and Group A in List 2) were more likely to state a higher number of items being correct. This is consistent with tax evasion being quite prevalent among respondents.

ence in the size of the treatment effects between the first and second list experiments, which provides strong evidence that these effects are not simply a matter of chance. Moreover, respondent in groups A and B each selected around 2.86 items on average across both list experiments (see row 3 of Table 2), which provides considerable reassurance that what is included as non-sensitive items in each of the specific list experiments is not affecting the results. In addition, there were no differences in the background characteristics of firms across the two groups (see Table A1 in the Appendix). The only meaningful difference between groups A and B was whether they were randomly allocated to receive the tax evasion item in the first or second list experiment.

Table 2: Main Results

	Control	Treatment	Difference	SE	N
List 1	1.161	1.435	0.274***	0.044	2330
List 2	1.465	1.715	0.250***	0.041	2322
List 1+2	1.316	1.571	0.255***	0.030	4652

Note: The first two rows of this table show the average number of true statements reported by respondents in groups A and B, for the first and second list experiments, as well as the differences between these two groups. The third row of the table shows pools both experiments together. Standard errors are presented in parentheses. *** corresponds with a p-value below 0.01.

Several robustness checks were conducted to illustrate the reliability of the main results. Firstly, we reproduce the results by applying weights to the sample of businesses to generate the weighted average treatment effects. The weights employed were based on firm size, sector, and region. This did not have a qualitative impact on the findings (see Table A2 in the Appendix), but as expected, the level of variance increases once weights are applied. Table A2 is restricted to all respondents who participated in both listing experiments. This shrinks the sample size to 2,272 but does not meaningfully change our results. Finally, we reproduce our analysis, including all respondents who refused to answer these questions and treated them as if they did not evade taxes. This still shows that substantial evasion is likely to be present.

5.2 Heterogeneous Effects

The double list experiment was included as part of the WBES which collects extensive information on firm characteristics and the beliefs of top managers or owners, which means that there are many dimensions along which heterogeneity in the main findings could be explored. We rely on a machine learning algorithm to identify the dimensions where substantial and consistent heterogeneity exists. Specifically, we use a causal forest technique[10] to identify where the greatest heterogeneity occurs across 27 potential dimensions that are captured in the survey prior to the double list experiment questions. A list of these 27 dimensions together with a measure of their relative heterogeneity in tax evasion prevalence is presented in Table A3 in the Appendix. We use this inductive approach to identify heterogeneity as ex ante, it is not immediately clear which dimensions will exhibit the most significant variation in tax evasion. Importantly, we focus our analysis on the dimensions where substantial and consistent heterogeneity exists in both list experiments. There are instances where differences across some dimensions are meaningful in only one of the list experiments and where the direction of heterogeneity varies. As we are interested in the most robust results, we do not focus on these dimensions where heterogeneity is not substantial or consistent across both list experiments.

[10]The causal forest function in the R package grf.